DOCUMENT RESUME

ED 220 519 TM 820 537

AUTHOR	Benson, Jeri; And Others
TITLE	Effects of Item Phrasing on the Factorial Invariance of Attitude Measures in Elementary School Children.
PUB DATE	Mar 82
NOTE	22p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Attitude Measures; Factor Analysis; *Factor Structure; Intermediate Grades; Item Analysis; *Response Style (Tests); Student Reaction; *Test Format; Test Interpretation; *Test Validity
IDENTIFIERS	*Likert Scales; *Phrasing

ABSTRACT

An empirical comparison of the validity of three Likert-type attitude scales measuring fourth to sixth grade students' attitudes toward school integration was conducted. The content-parallel forms differed only in terms of item phrasing and were composed of either all positive, all negative or a mix of items. Different factor solutions underlying the scales were hypothesized for the All-Positive, All-Negative and Mixed Forms. The different structures were compared statistically using the LISREL program to determine whether elementary children respond differently to positive and negative phrasing and to examine the utility of alternating item phrasing to enhance validity. Results indicated that the subjects responded differently to the same phrase stated positively and negatively, implying that the technique of balancing item phrasing to control response sets is not effective for younger children. (Author/CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED220519

Effects of Item Phrasing on the Factorial Invariance of Attitude Measures in Elementary School Children

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Jeri Benson
Dennis Hocevar
Cynthia Cohen

University of Southern California

The purpose of the study was to empirically compare the validity of three Likert-type attitude scales using 4-6 grade children. The three scales were composed of either all positive, all negative, or a mix of positive and negative items. Different factor solutions were hypothesized to underlie the scale. The different structures were then compared statistically using the LISREL program. Results indicated that elementary students responded differently to the same positively and negatively phrased items. Thus, the technique of balancing item phrasing to control response sets does not appear to hold true for younger children.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. Benson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 537

A paper presented at the annual meeting of
the American Educational Research Association
New York, March 1982

Effects of Item Phrasing on the Factorial Invariance of Attitude Measures in Elementary School Children

Jeri Benson
Dennis Hocevar
Cynthia Cohen

University of Southern California

Anderson (1981) has argued quite effectively as to the importance of including affective variables in the elementary curriculum. He emphasized that affective characteristics facilitate desired cognitive goals of the schooling process and are themselves desired goals of the schooling process. Similarly, Bloom (1978) stated that schools should produce "independent learners" who are able to engage in higher level thinking, develop confidence in their abilities, are intrinsically motivated and possess a degree of social responsibility. Thus, with more emphasis being placed on the affective aspect of the elementary curriculum the need to assess the effectiveness of these interventions will increase.

Most affective interventions are currently assessed via attitude surveys, however, the development of scales to assess elementary student attitudes poses very complex methodological problems for the test constructor. Definitions of the attitude to be measured and establishing the reliability and validity are difficult topics to deal with since the in-

terests and attitudes of young children change quickly, reading skills vary and the attention span of the elementary child is limited. In addition, the question of whether a response set might influence the results and hence the validity of an attitude survey for children has not been addressed directly in the literature. The major concern of the present study is the influence of response sets on the measurement of children's attitudes.

The possibility of a response set influencing the validity of cognitive instruments was first reported by Cronbach (1946; 1950). He defined response set as "any tendency causing a person consistently to make different responses to test items than he would have had the same content been presented in a different form" (1946, p.476). Rorer (1965) after reviewing the literature on response sets in cognitive and noncognitive instruments concluded that an individual's response set can be influenced by item content and may account for a significant proportion of response variance.

In order to reduce the possibility of a response set operating in affective instruments, it has been suggested to use a forced choice format or to use an equal number of positive to negative statements (Anderson, 1981; Mehrens & Lehmann, 1980). Presently, most attitude scales are developed to include both positively and negatively phrased items to reduce the possibility of a response set occuring. However, the recommendation to use an equal number of positive and negative items is based upon research with adults, most often college students, and not elementary school students.

In addition this recommendation is based on the assumption that positive and negative test items measure the same construct. If this were not the case, the instrument would be less reliable when negative and positive items are combined into the same index. Finally, this recommendation assumes that combined positive and negative items yields a more valid index. While there is a wide acceptance of this assumption, there is little research on its tenability.

The purpose of the present study was to examine the effects of positive and negative item phrasing on elementary school students' responses to three variations of Likert type attitude scales. The three variations were composed of either all positive, all negative or a mix of positive and negative items. Since evaluators and researchers are increasingly using affective instruments to gather data on the effects of curricular programs that impact on elementary students, it is encumbant upon psychometricians to carefully study how the above suggestions may or may not be an aid to reducing response set in attitude scales administered to children.

METHOD

Sample

The sample consisted of all fourth, fifth, and sixth grade students (N=622) from three schools who were part of a district-wide court-ordered integration plan. The data were collected as a part of an assessment of an integration study requested by the three schools.

Instrumentation

Three content parallel Likert forms were developed to assess student attitudes toward integration. The forms differed only in terms of item phrasing. The first form contained 15 positively phrased items. The second form contained the same 15 items negatively phrased. The third form contained the same 15 items, eight were positively phrased and seven were negatively phrased. All items were presented in the same sequence. Items written for the three forms were developed based upon objectives the schools proposed during the first year of court-ordered integration. Thus, a student's attitude toward integration was defined by the responses to items developed to assess the school's objectives. The items were judged for content relevance by the persons responsible for conducting the integration program at each school. An example of an item from two of the forms is presented below:

Positive Form: I like going to an integrated school.
Negative Form: I do not like going to an integrated school.

The items appearing on the Mixed Form were randomly chosen from the Positive and Negative Forms. The coefficient alpha reliability of the three forms a well as the reliability estimates for each grade level by form is shown on Table 1.

Insert Table 1 about here

Procedures

Each student responded anonymously to one of the three forms which were randomly administered in each class to all students by their classroom teacher at the end of the Spring semester. Each item was scored by weighting the response on a Likert scale where 1 = strongly agree and 5 = strongly disagree.

Initially, a form by grade (3x3) analysis of variance (ANOVA) was run on the dependent variable of student attitude to determine if there were any mean differences between forms at each grade level. A statistically significant interaction (F (4,591)= 3.54, $p<.05$) was found. Subsequent post hoc comparisons were run and the following results were noted. The fourth grade students responded more favorably toward integration on the Positive Form than on the Negative or Mixed Forms ($\bar{X}$ = 35.44, $\bar{X}$ = 39.43, $\bar{X}$ = 40.66 respectively). The sixth grade students responded more favorably toward integration on the Negatively phrased Form than did either the fourth or fifth grade students ($\bar{X}$ = 34.95, $\bar{X}$ = 40.66, $\bar{X}$ = 38.57 respectively). Thus, there appeared to be a differential response pattern for the elementary children across the forms.

Insert Figure 1 here

Since the preliminary results reported above indicated a potential validity problem in that mean differences appeared to depend upon form, an investigation of the factor structure underlying each form was undertaken. For the All Positive and All Negative Form it was hypothesized that a general trait-one factor solution would best fit the data. For the Mixed Form it was hypothesized that a two factor solution would best fit the data. The two factors would be due to item phrasing, one factor composed of the positive items and one factor composed of the negative items.

In order to test the above hypotheses the following models were proposed and examined using the LISREL IV program (Jöreskog & Sörbom, 1978):

a) Model 1 - a one factor solution holding the factor loadings invariant;

b) Model 2 - a one factor solution allowing the factor loadings to be unrestricted across each form;

c) Model 3 - an unrestricted one factor solution for the Positive and Negative Forms and an unrestricted two factor solution for the Mixed Form.

DATA ANALYSIS

The responses to each form were converted to separate covariance matrices and were used as input to the LISREL program to test each model. Models 1 and 2 were used to examine whether each form had the same or a different factor structure, and Model 3 was used to directly test the hy-

potheses of interest. The LISREL program requires that three matrices be specified prior to analysis. The lambda pattern matrix allows the user to identify which variables are free and thought to load on a given factor and which variables are fixed at zero and thought not to load on a factor. In this way, the hypothesized factor pattern described in Models 1 - 3 can be established. The psi matrix allows one to see if the factors are correlated by setting them free or uncorrelated. Finally, the theta epsilon matrix which provides an estimate of the error or uniqueness of each variable must be specified. In the present analysis all theta estimates were left free to be estimated. To test for Model 1, the three groups were analyzed simultaneously by specifying an invariant solution for the lambda matrix at the beginning of the program. This procedure was used to determine whether the factor structure was the same for all forms. The lambda pattern matrix and the theta epsilon matrix were set to be free and the psi matrix was fixed since only one factor was being estimated. For Model 2, the same program as Model 1 was run; however, the lambda matrix was not set for an invariant solution but for a free or unrestricted solution. In this way the items on each of the three forms were free to take on the loadings that best represented the data. For Model 3, the lambda and theta epsilon matrices were set to specify a one factor model for the Positive and Negative Forms and a two factor model for the Mixed Form. The psi matrix for the two factor model was set free to determine to what extent the two factors would be correlated.

The LISREL procedures provide a chi-square statistic as well as estimates of the factor and item uniquenesses to indicate model-data fit. The results were used to determine whether elementary children respond differentially to positive and negative item phrasing and hence examine the utility of alternating item phrasing to enhance validity.

RESULTS AND DISCUSSION

The chi-square test of model-data fit is reported on Table 2. The lower the chi-square value the better the model-data fit. Thus, a non-significant chi-square value is desired as it would indicate the model proposed fit the data (Bentler & Bonett, 1980). Since the covariance matrix was used in obtaining the theta estimates, the values can exceed one and should be interpreted relative to each other.

Insert Table 2 about here

Model 1, where the hypothesis of the same factor loadings occuring across each form was tested, revealed a significant $\chi_1^2 = 845.3562$ with 300 degrees of freedom ($p < .001$). This indicated that the covariance matrix generated by the one factor solution was significantly <u>different</u> from the original covariance matrix used as input. The chi-square values for Models 2 and 3 were also significant at $p < .001$. Since the statistical significance of the chi-square test can be influenced by sample size, Bentler and Bonett (1980)

recommend the comparison of alternate models to determine the relative effectiveness of one model over another in terms of fit. Bentler and Bonett suggest the following chi-square difference test be used to test the significance of model-data fit:

$\chi^2_{1-2} = \chi^2_1 - \chi^2_2$ with degrees of freedom = $df_1 - df_2$, (1)

where χ^2_1 represents one hypothesized model and χ^2_2 represents a different hypothesized model with their corresponding difference in degrees of freedom.

Since three models were hypothesized in the present study, formula 1 can be applied to make comparisons between each of the observed chi-square values reported on Table 2 to determine if Model 2 was significantly better than Model 1 and if Model 3 was significantly better than Models 1 and 2. When Models 1 and 2 were compared using formula 1, the result (χ^2_{1-2}=190.82 and 30 degrees of freedom) was statistically significant $p<.001$. The implication being that Model 2 (which tested whether the factor structure differed across the three forms) was a better model to represent the data than Model 1 (which tested whether the factor structure was the same across the three forms). Therefore, the fact that Model 2 fit the data better than Model 1 provided additional evidence in support of the ANOVA results reported earlier where a differential response pattern across forms was indicated by the significant interaction between form and grade level. Bentler and Bonett (1980) developed an additional test to determine the usefulness of alternative models in terms of incremental fit. An incremental index of fit pro-

vides information regarding the practical significance of the alternative models independent of sample size. The fit index is defined as follows:

$$\Delta_{12} = \chi_1^2 - \chi_2^2 \; / \; \chi_1^2 \quad , \qquad (2)$$

where χ_1^2 and χ_2^2 reference the chi-square values for alternative models thought to underlie the data. Bentler and Bonett (1980) suggest that the most restrictive model be used in the denominator of Delta. Since Model 1 in the present study was the most restrictive, (factor structure invariant across the three forms) it was placed in the denominator. For the Model 1 - Model 2 comparison of fit, Δ_{12} was .226 (see Table 3). The values of Delta range between zero and one and as such Model 2 represented an improvement in fit over Model 1.

Insert Table 3 here

Further evidence of differential factor structure was provided by noting on Table 4 the lambda and theta estimates obtained for Model 2.

Insert Table 4 about here

All of the factor loadings for the Positive Form were statistically significant $p < .05$ (except item 6) as were all of the factor loadings for the Negative Form $p < .05$ (except item 8). The data indicated strong support for a one factor model for the Positive and Negative Forms. However,

only nine of the fifteen items on the Mixed Form were statistically significant (items 2,7,10,4,6,9,11,12,13). Glancing at the loadings reported on Table 4 for the Mixed Form, one notices a very distinct pattern in that the positively phrased items (1,2,5,7,8,10,14,15) have rather low loadings whereas the negatively phrased items (3,4,6,9,11,12,13) on the whole have much higher loadings. This pattern suggests a two factor structure might better represent the data obtained from the Mixed Form.

When Models 1 and 3 were tested using formula 1 the result (χ^2_{1-3} = 252.81 and 31 degrees of freedom) was statistically significant at $p < .001$. The increment of fit of Model 3 over Model 1 (Δ_{13} using formula 2) was .300. Thus, model 3 not only provided a statistically better fit to the data, but independent of sample size Model 3 provided a much better description of the data than Model 1 in terms of incremental fit. Also, when Models 2 and 3 were tested using formula 1 the result (χ^2_{2-3} = 61.99 and 1 degree of freedom) was again statistically significant at $p < .001$. The increment of fit of Model 3 over Model 2 was Δ_{23} = .095. Again, Model 3 was a better fit to the data than Model 2. In addition, the lambda and theta estimates shown on Table 5 indicated that for the Mixed Form a two factor model was a better fit of the data than the one factor model (shown on Table 4) as all of the item loadings for the two factor model were statistically significant at $p < .05$ (except item 3).

Insert Table 5 about here

Further support for the two factor model for the Mixed Form was illustrated by the correlation between the two factors (psi matrix) where r = .114, an indication that the two factors are indeed distinct.

The data from the present study have provided strong evidence that alternating item phrasing in an effort to diminish the possibility of a response set occuring resulted in a differential response pattern for the elementary children. The significant interaction between grade and form from the ANOVA analyses along with Figure 1 and the discrepant reliabilities across the forms (Table 1) provided an initial indication that the children were not responding to the same items in a like manner. Noting the mean ratings illustrated in Figure 1, the fourth and fifth grade students tended to show a more negative attitude on the all Negative Form and the fourth graders also tended to show a more negative attitude on the Mixed Form. Subsequent analysis using the LISREL IV program to statistically compare three proposed models thought to underlie the data were tested. Model 1 was the most restrictive where a one factor solution was hypothesized to represent the data and that the solution would be invariant across the Positive, Negative and Mixed Forms. Model 2 was a one factor solution where the structure for each of the three forms was not restricted so that the factor loadings were free to conform to each set of data. Mo-

del 3 was the model of interest in that it sought to explain the differential response pattern noted in the ANOVA and reliability results reported previously. Model 3 proposed a one factor unrestricted solution for the Positive and Negative Forms since the item phrasing was constant for these forms and a two factor unrestricted solution for the Mixed Form since it was hypothesized that alternating the item phrasing within the scale would result in a method effect. That is, the child's response pattern was more dependent upon the phrasing of the item than the content of the item. As the chi-square tests of significance and Delta tests of incremental fit have revealed, Model 3 provided a better match to the data than Models 1 and 2.

Since much of the research regarding alternating item phrasing within a cognitive or affective instrument was based upon adult responses it is important to examine whether the same conclusion would be made when the respondents were young children. The ANOVA and LISREL results of the present study indicated that elementary students responded differently to the same positively and negatively phrased items. Hence, the technique of balancing item phrasing in order to increase the reliability and validity of score interpretations from attitude surveys does not appear to hold true for younger children.

Most affective measurement in the schools is the self-report variety (quite similar to the Likert attitudinal instrument used in the present study) therefore, psychometricians and researchers in general need to be aware of the

possible effect of alternating item phrasing will have on the assessment of attitudes in children. The general recommendation to use an equal number of positively and negatively phrased items may do more harm to a research study or evaluation than it was originally thought to correct, especially when the respondents are younger students. Therefore, one ought to carefully consider whether the phrasing of the items in an attitude survey for younger children might in some way effect their response, and thus, bring into question the validity of the results of the attitude survey. Based upon the results of the present study it is recommended that attitude surveys developed for elementary children use all positively phrased items so that the expressed attitude is not confounded by the phrasing of the item. The use of all negative phrasing is not recommended since it was shown that younger children tended to indicate a more negative attitude under that condition. However, it is recognized that additional research certainly needs to be conducted with the elementary student population using different attitude scales to verify the present findings and recommendations.

REFERENCES

Anderson, L.W. Affective characteristics in the schools. Boston: Allyn and Bacon, Inc., 1981.

Bentler, P.M., & Bonett, D.G. Significance tests & goodness of fit in the analysis of covariance structures. Psychological Bulletin, 1980, 88, 588-606.

Bloom, B.S. New learner: Implicatons for instruction and curriculum. Educational Leadership, 1978, 35, 553-576.

Cronbach, L.J. Response sets and test validity. Educational and Psychological Measurement, 1946, 6, 616-623.

Cronbach, L.J. Further evidence on response sets and test design. Educational and Psychological Measurement, 1950, 10, 30-31.

Jöreskog, K.G., & Sörbom, D. LISREL IV: Analysis of linear structural relationships by the method of maximum likelihood. Chicago: International Educational Services, 1978.

Mehrens, W., & Lehmann, I. Standardized tests in education, (third edition). New York: Holt, Rhinehart & Winston, 1980.

Rorer, L.G. Psychological Bulletin, 1965, 63, 129-156.

Table 1

Coefficients Alpha Reliability Estimates
By Grade Level and Form

Grade	Form		
	Positive	Negative	Mixed
4	.72(69)[a]	.76(62)	.61(70)
5	.80(74)	.77(70)	.66(76)
6	.77(62)	.80(58)	.65(53)
Total	.78(212)	.78(201)	.65(203)

[a] The sample size for each grade level by form is shown in parentheses. For each form some students failed to code their grade level thus, the total number of students responding to each form is slightly higher than the sum of each grade level.

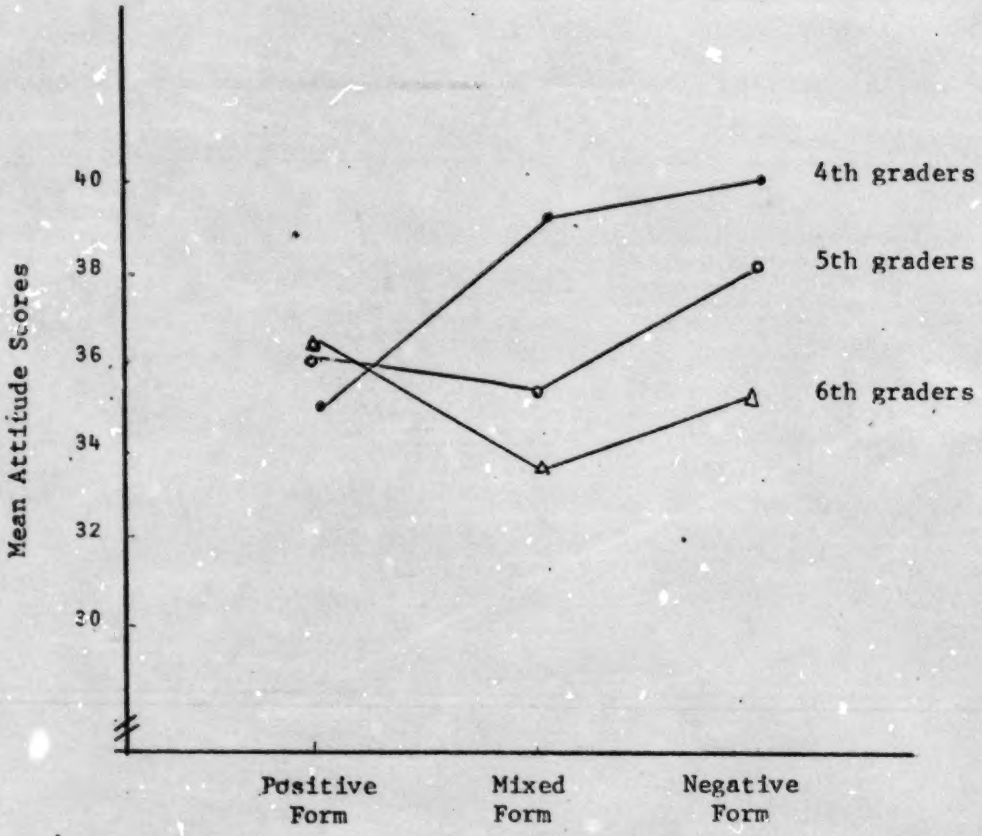

Figure 1. Mean Attitude Scores Across Forms and Grade Level

Note. The responses were keyed such that the lower score represented a more favorable attitude.

Table 2

Models Tested Regarding the Factor Structure of The Three Forms

Model	Description	X^2	df
1	Same Factor Structure across forms (invariant)	845.3562	300
2	Different Factor across forms (unrestricted)	654.5395	270
3	One Factor Underlying Positive & Negative Form Two Factors Underlying Mixed Form	592.5472	269

Table 3

Delta Coefficients[a] for Model Comparisons

	Model 1	Model 2	Model 3
Model 1	--	.226	.300
Model 2		--	.095
Model 3			--

[a] Delta Coefficients range from 0 to 1

Table 4

Lambda and Theta Estimates[a]

For Model 2

	Positive Form		Mixed Form		Negative Form	
Item	Factor 1	Theta	Factor 1	Theta	Factor 1	Theta
1	.425	.602	.046	.624	.646	.928
2	.393	1.194	-.190	1.520	.484	1.445
5	.338	.609	.027	.928	.819	1.289
7	.643	2.259	.247	2.093	.312	2.159
8	.534	1.891	.076	2.136	.099	2.154
10	.393	1.235	.188	1.301	.407	1.154
14	.510	1.558	-.166	1.528	.298	1.851
15	.657	1.191	.169	1.549	.697	1.412
3	.662	1.624	.221	2.147	.538	1.788
4	.657	.913	.654	1.155	.551	1.417
6	.098	2.018	.648	1.317	.563	1.647
9	.630	.708	1.00	.759	.743	.928
11	.796	.619	.791	.701	.641	.790
12	.642	.989	.735	1.109	.699	1.223
13	.660	.843	.799	.955	.874	.989

Note: For the Mixed Form items 1, 2, 5, 7, 8, 10, 14, 15 were positively phrased and items 3, 4, 6, 9, 11, 12, 13 were negatively phrased.

[a] The Lambda and Theta values were set free to best represent the data (for each form)

Table 5

Lambda and Theta Estimates[a]

Two Factor - Mixed Form

	Factor 1	Factor 2	Theta
1	.220	.0	.577
2	.347	.0	1.436
5	.397	.0	.771
7	.372	.0	2.016
8	.485	.0	1.907
10	.487	.0	1.099
14	.429	.0	1.372
15	.638	.0	1.170
3	.0	.218	2.148
4	.0	.659	1.148
6	.0	.637	1.331
9	.0	1.000	.744
11	.0	.788	.705
12	.0	.738	1.105
13	.0	.796	.960

[a] The Lambda and Theta values were set free to best represent the data.

Note: For the Mixed Form items 1, 2, 5, 7, 8, 10, 14, 15 were positively phrased and items 3, 4, 6, 9, 11, 12, 13, were negatively phrased.